International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

Robust Estimation in Stratification Sampling

Ajiteru S. Oyeyemi
ICT Centre,
Federal Polytechnic Offa,
Offa, Kwara State, Nigeria

***Abstract*:- The estimation of any variable of interest such as the one considered in this study: monthly allowance and expenditure of students in the university depends on the sampling scheme used. In this research, estimation using simple random sampling and stratification sampling is considered. Statistical Package for Social Sciences (SPSS) was used for easier computation of the estimates needed. It was found that stratification sampling scheme gives a better variance and it is therefore recommended.**

Keywords: Estimation, Variable, population, Sampling and Stratification

INTRODUCTION

Sample design has two aspects: a selection process, the rules and operations by which some members of the population are included in the sample; and an estimation process (or estimator) for computing the sample statistics, which are sample estimates of population values (Hidiroglou and Rao,1983).

Kish (1965) opined that survey objectives should determine the sample design; but the determination is actually a two-way process, because the problems of sample design often influence and change the survey objectives. We shall encounter examples of the ways in which survey objectives and sample design interact to produce overall survey designs.

A dialogue between the researcher and the sampler must occur before any aspect of survey design is "frozen," because a change in one aspect may dictate a change in others. Instead of a dialogue, the decisions may involve a larger cast: sampler, researcher, and consumer; and the last, perhaps the grantor of the project, may feel behind him the silent pressure of the "ultimate consumers" of the data – the members of a profession or, perhaps, a wider public. The dialogue may occur silently within one head, if the researcher and sampler are one; but the dialogue should nevertheless take place (Kuk, 1988 and Okafor, 2002).

Most samples are prepared by statistician and other researchers who are not primarily sampling specialists. Nevertheless, it is helpful, although sometimes difficult, to separate sampling design from the related activities involved in survey research. The sample design covers the tasks of selection and estimation for making inference from sample value to the population value. Beyond this are the problems of making inferences from the survey population to another and generally broader population, with measurements free from error (Rao, 1994).

AIM AND OBJECTIVES

The aim of this study is to compare variances in linear combination using simple random sampling and stratification sampling.

The objectives are to:

(i) Estimate the mean monthly income and expenses of students in the university

(ii) Present a better procedure for estimation when the mean monthly incomes and expenses of students in the university are involved.

Characteristics of population elements are transformed to variables Y_i by the survey operations of measurement. Some literature deals directly with the statistical populations of the variables Y_i. But I prefer to say that the ith element has the variable Y_i. This permits us to talk of the many variables (Y_i, X_i, Z_i, W_i, P_i and so on) of the same element (Rao, Kovar and Mantel, 1990). We can also consider relationships between variables of an element, changes of variables, and accuracy of measurements of variables. A statistic based on the variables found in a sample results in a random variable is what we call a variate (Kendall and Buckland, 1957).

METHODOLOGY

Y_i = value of the Y_i variable for the jth sample element

$$y = \sum_j^n y_j$$

$$\bar{y} = \frac{1}{n}\sum_j^n y_j \qquad (1)$$

$$s_y^2 = \frac{1}{n-1}\sum_j^n (y_j - \bar{y})^2$$

$$Var(\bar{y}) = E\{(\sum_j^n y_j - n\bar{Y})^2\} = E\{(\Sigma_{j,k}^n)(y_j - \bar{Y}) + \sum_j^n (y_j - \bar{Y})(y_k - \bar{Y})\}$$

$$Var(\bar{y}) = \frac{n}{N}\sum_i^N \{Y_i - \bar{Y}_i\}^2 + \frac{n(n-1)}{N(N-1)}[\sum_{i\neq h}^N Y_i - \bar{Y}\{Y_i - \bar{Y}\}] = n\sigma_y^2 + \frac{n(n-1)}{N(N-1)}[\{\Sigma_i^N (Y_i - \bar{Y})\}^2 - \Sigma_i^N (Y_i - \bar{Y})^2] = (1-f)\, nS_y^2$$

(2)

where f = n/N

Also,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

$$\text{Var}(N\bar{y}) = N^2 \text{Var}(\bar{y}) = \frac{1-f}{n} S_y^2 \quad (3)$$

MEAN AND VARIANCE OF SIMPLE RANDOM SAMPLING

The simple mean of the sample of a selection is the SRS mean, and we distinguish it with the subscript 0:

$$\bar{y}_0 = \frac{1}{n} \sum_j^n y_j = \frac{1}{n} [\, y_1 + y_2 + \ldots + y_n \,] \quad (4)$$

The results of an SRS selection may be used for other estimators also, for example, with post-stratification or with a ratio estimator (Sarndal and Wright, 1992). But we treat those separately as other designs. Simple random sampling is a sample design specifying both the SRS selection and the simple mean estimate. The variance of the SRS mean $\bar{y}_0$ is computed as

$$\text{var}(\bar{y}_0) = (1 - f) \frac{S^2}{n}$$

where

$$s^2 = \frac{1}{n-1} \sum_j^n (y_j - \bar{y})^2$$

$$= n \frac{\Sigma^n y_j{}^2 - y^2}{n(n-1)}$$

The standard error of $\bar{y}_0$ is the root of its variance:

$$\text{se}(\bar{y}_0) = \sqrt{\text{var}(\bar{y}_0)}$$

$$= \sqrt{1-f}\, \frac{s}{\sqrt{n}} \quad (5)$$

Sometimes we may want to estimate $Y = N\bar{Y}$, the aggregate or total of the Y_i variable in the population. A simple estimator of Y is the $N\bar{y}_0$ and its standard error is estimated by

$$\text{se}(N\bar{y}_0) = N \sqrt{1-f}\, \frac{s}{\sqrt{n}} \quad (6)$$

We can also point out that the expected value of s^2 in SRS is

$$E(s^2) = \frac{N}{N-1} \sigma^2 \quad (7)$$

This is shown as the expected value of the sample estimate of the variance of the mean is

$$E\left(\frac{1-f}{n} s^2\right) = \frac{N-n}{N-1} \frac{\sigma}{n} \quad (8)$$

For the difference $(\bar{y} - \bar{x})$ of two means, the variance is simply the sum of the two variances if the two samples are independent (Thompson, 1992). But if the two means are not independent, a covariance term must be subtracted from the sum of the variances: $\text{var}(\bar{y} - \bar{x}) = \text{var}(\bar{x}) + \text{var}(\bar{y}) - 2\,\text{cov}(\bar{y}, \bar{x})$.

For n pairs of values, each pair selected with SRS, the difference has the variance:

$$\text{var}(\bar{y} - \bar{x}) = \frac{1-f}{n} (S_x^2 + S_y^2 + 25_{yx}) \quad (9)$$

Note also the use of covariance of the two variables. The statistics resembles the variance. But contains cross-product terms instead of the squared terms of the variance

$$\text{cov}(\bar{y}_0, \bar{x}_0) = (1 - f) \frac{S_{yx}}{n}, \; S_{yx} = \frac{1}{n-1} \left[\sum_j^n y_j x_j - \frac{y_x}{n} \right] \quad (10)$$

Note also that for the pairs of elements

$$(\bar{y} - \bar{x}) = \frac{\Sigma_j^n y_j}{n} - \frac{\Sigma_j^n x_j}{n}$$

$$= \sum_j^n \frac{d_j}{n}$$

Hence we may treat this as the mean of a sample of n element $(y_j - x_j) = d_j$. The variance can also be computed as

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

$$\text{var}\left(\frac{1}{n}\sum_j^n d_j\right) = \frac{1-f}{n} s_d^2 \qquad (11)$$

where

$$s_d^2 = \frac{1}{n-1}\left[\sum_j^n d_j^2 - \frac{\Sigma d_j}{n}\right],$$

which numerically equal to (15). The covariance is absent for two independent samples, but present for two overlapping samples. The variance of the difference becomes more complicated if the two samples are neither completely independent nor completely overlapping.

The subclass mean $\bar{y}_m = \sum_j^m y_j / m$ from an SRS of n elements can be treated as an SRS of m elements. That is, we consider the variance of the sample conditional on obtaining a sample of m elements:

$$(\bar{y}_m) = \frac{1-f}{m(m-1)}\sum_j^n (y_j - \bar{y}_m)^2 = \frac{1-f}{m} s_m^2 \qquad (12)$$

We can use f = m/M if we know M, the population size of the subclass. If we do not know M, we can use f = n/M, neglecting the difference form m/M. But if we want to estimate Y = MY, the population total subclass, then knowledge of the subclass size M becomes important. If M is known, then (M $\bar{y}_m$) has the variance M^2 var ($\bar{y}_m$). If we do not know M and use (mN/n) $\bar{y}_m$ to estimate Y, then the element variance $s_{\bar{y}}^2$ is increased to $[s_m^2 + (1 - \bar{m}) \bar{y} m^2]$.

The formula with the possible exception of the factor (1 - f) = (1 - n/N). This factor is usually called the finite population correction, briefly fpc. When sampling without replacement, it appears as a correction factor to the main portion of the variance terms, which is s^2 / n for SRS. If we think of a fixed sample size n being applied to larger and larger populations, the sampling fraction g = n/N tends to zero, and the factor (1 – f) approaches 1. Multiplication by one has no effect, and the fpc can be omitted when the population is much larger than the sample. For an "infinite population" the factor disappears from the variance formula; hence, its name. Also, when selecting with replacement, the factor (1 – f) becomes 1 and disappears. The effect is similar to selection from an infinite population.

The sampling fraction is usually small, because the population is large. The aims of research generally concern inferences from about large populations or confined to a small population. This often is hopefully considered a "sample" for making inference about some much larger actual population or theoretical universe. But census aimed specifically as small populations do occur, and sometimes these run into larger fractions of 10 percent and more. In these rare cases the fpc is needed. Note that the variance can be written as var ($\bar{y}$) = S^2 / n' where n' = n/(1 – n/N) = nN/(N – n). From this we easily note that n = n'/(1 + n'/N). In the words, the effect of (1 – n/N) is to increase the "effective sample size" from n to n'. It might be convenient to write all the variance formulas with this convention.

RELATIVE ERROR

In some situation it is useful to consider some relative measures instead of the absolute measures of the variation. The absolute measures, the standard deviation and the standard error, appear in the units of measurement of the variable, and this causes difficulties in some comparisons. Common relative measures are the coefficients of variation, in which the unit of measurement is cancelled by dividing the mean. The element coefficient of variation is derived from the standard deviation:

$$C_y = \frac{S_y}{\bar{Y}}, \text{ estimated by } c_y = \frac{S_y}{\bar{y}} \qquad (13)$$

The coefficient of variation of the mean ($\bar{y}$) is derived similarly from the standard error:

$$CV(\bar{y}) = \frac{SE(\bar{y})}{\bar{Y}}, \text{ estimated by } cv(\bar{y}) = \frac{se(\bar{y})}{\bar{y}} \qquad (14)$$

The squares of these quantities correspond, respectively, to the variances of the element and of the mean:

$$C_y^2 = \frac{S_y^2}{Y^2} \text{ estimated by } c_y^2 = \frac{S_y^2}{\bar{y}^2}$$

is an element relative variance and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

$CV^2(\bar{y}) = \frac{Var(\bar{y})}{\bar{Y}^2}$ estimated by $cv^2 = \frac{var(\bar{y})}{\bar{y}^2}$

is a relative variance of the mean ($\bar{y}$).

Coefficients of variation are useful for variables that are always or mostly positive; these occur frequently in surveys, especially as "count data". Comparison of the variability of these items often becomes more meaningful when expressed in relative terms. For example, in comparing the "income spread" in two countries, the use of the two standard deviations would be confused by the different monetary units as well as by different standard of living; but coefficients of variation may provide a reasonable comparison in term of average income.

$cv(N\bar{y}) = cv(\bar{y})$ (15)

The general expression holds for different sample designs. Specifically, for SRS samples we can use

$cv(\bar{y}_0) = cv(N\bar{y}_0) = \sqrt{1-f}\ \frac{s_y}{\bar{y}_0\sqrt{n}}$ (16)

In some situations the coefficients of variation should be used only with caution, or not at all for the following reasons: (1) If the mean of the variable is close to zero, the coefficients of variances are large and unusable and (2) For binomial variable, the element variance is the same P (P – 1) for both P and 1 – P; but the coefficients of variation differ, depending on the arbitrary decision of which side of the binomial but is regarded as P and which as Q. That is:

$$C_y^2 = \frac{P(1-P)}{P}$$

and

$$cv^2(p) = (1-f)\frac{(1-P)}{P(n-1)} \quad (17)$$

The element relative variance $C_v^2 = (1 – P)/P = 1$ for P = 0.05. It increases rapidly for small values of P.

LINEAR COMBINATIONS USING STRATIFICATION SAMPLING

Mean of Linear Combinations Using Stratification Sampling

In stratified sampling where the population of N units is first divided into subpopulations of N_1, N_2, ..., N_L units respectively. These subpopulations are non overlapping and together they comprise the whole of the population, so that

$N_1, N_2, ..., N_L = N$

Then;

$\bar{y}_{lst} = \sum_{h=1}^{L} W_h \bar{y}_h$ is a linear function of the $\bar{y}_h$ with fixed weights W_h. (18)

Variances for Linear Combinations Using Stratification Sampling

We obtain variances for some more complicated linear combinations that we shall need later. The sum of H random variables, weighted by the constant factors $W_h$has the variance:

$Var\{\Sigma W_h y_h\} = \Sigma_h W_h^2\ Var(y_h) + 2\Sigma_{h<y} W_h W_g\ Cov(W_h y_h)$ (19)

A common example is the sum or difference of two random variables y_1 and y_2, when $W_1 = 1$ and W_2 is either 1 or -1:
$Var(y_1 \pm y_2) = Var(y_1) + Var(y_2) \pm 2Cov(y_1, y_2)$
The covariance vanishes if y_1 and y_2 are uncorrelated. Another important special case when all H variate are uncorrelated, because they are based on independent samples from H strata. Then all the covariance vanish and
$Var\{\Sigma W_h y_h\} = \Sigma W_h^2\ Var(y_h)$ (20)
We can consider the covariance of the sums $\Sigma W_h y_h$ and $\Sigma V_h x_h$ of two sets of random variables, again assuming independence between the H sets; for example, these could be pairs of measurements on H independently selected elements:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

$$\text{Cov}\{\Sigma W_h y_h , \Sigma V_h x_h\} = \Sigma W_h y_h , \text{Cov}(y_h , x_h) \quad (21)$$

When constants W_h and y_h are all unity, we have

$$\text{Cov}\{\Sigma y_h , \Sigma x_h\} = \Sigma \text{Cov}(y_h , x_h) \quad (22)$$

The formulas for variances and covariance of linear combinations were developed for population values, written with capital letters as Var and Cov. But they apply also to their sample estimates, which we write with lower case letters as var and cov. Summation of the estimated variances and covariances for sample totals within strata is simple and frequently needed. Therefore, we employ the brief notation dy_h^2, dx_h^2 and $dy_h dx_h$. When the y_h and x_h represent two variates for selections that are independent between strata, we have

$$\text{Var}\{\Sigma y_h\} = \Sigma \text{var}(y_h) = \Sigma dy_h^2$$

$$\text{Var}\{\Sigma x_h\} = \Sigma dx_h^2$$

$$\text{Cov}\{\Sigma y_h x_h\} = \Sigma dy_h dx_h \quad (23)$$

RESULTS

Here, an estimation of mean monthly allowance of students in Mathematics and Statistics, Department of Federal University of Technology, Minna is considered. The set of data gathered is stratified into two strata: male (stratum 1) and female (stratum 2). We present the summary of results generated as follows for simple random sampling and stratification sampling in tables 1 and 2 respectively:

Table 1: Estimates for Simple Random Sampling

Estimate / Parameter	X	Y
n	38	38
Mean	19,750.00	18, 170.00
s^2	1,540,000,000.00	1,299,998,914.00
Standard Error of Mean	6,357.58	5,848.97

Table 2: Estimate for Stratification Sampling

Estimate / Parameter	X	Y
n	38	38
Mean	12,263.16	11,282.11
s^2	62,469,417.00	52,874,114.37
Standard Error of Mean	6,357.16	5,848.59

DISCUSSION OF RESULTS

From tables 1 and 2, it was found that the variance and the corresponding standard error in the case of stratification sampling is less than that of the variance and the corresponding standard error in the case of simple random sampling. For a simple random sampling, the means of x and y are ₦19,750.00 and ₦18,170.00 with standard errors of 6,357.58 and 5,848.97 respectively. That is the average monthly allowance for a student is ₦19,750.00 while the average monthly expenditure for a student is ₦18,170.00. Also, for a stratification sampling, the means of x and y are ₦12,263.16 and ₦11,282.11 with standard errors of 1,282.16 and 1,179.59 respectively. That is the average monthly allowance for a student is ₦12,263.16 while the monthly expenditure for a student is ₦11,282.11.

CONCLUSIONS AND RECOMMENDATIONS

From the findings above, it was observed that the estimation procedure using stratification sampling is better than linear combination of simple random sampling. An approach that is better in sampling technique is always being adopted when there is a need for computation involving such variable of interest. The study shows that the estimation from stratification sampling scheme gives the minimum variance and standard error of mean. Hence, estimation using stratification sampling scheme is recommended.

REFERENCES

[1] Hidiroglou, M. and Rao, J. (1983). On two sample schemes of unequal probability sampling without replacement. Journal of the Indian Statistical Association, 3: 173-180.
[2] Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of sampling without replacement from a finite universe. Journal of American Statistical Association. 47: 663-685.
[3] Kish, L. (1965). Survey Sampling. John Wiley and Sons.
[4] Kuk, A. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. Biometrika, 75: 97-103.
[5] Okafor, F. C. (2002). Sample Survey Theory with Applications, Afro-Orbis Publications, Nigeria.
[6] Rao, J. (1994). Estimating totals and distributions functions using auxiliary information in the estimation stage. Journal of Official Statistics, 10: 153-166.
[7] Rao, J., Kovar, J. and Mantel, H. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, 77: 365-375.
[8] Sarndal, C. E. and Wright, R. L. (1992): Design-based and model-based inference in survey sampling. Scandinavian Journal of Statistics, 5: 27-52.
[9] Thompson, S. K. (1992). Sampling. John Wiley and Sons, New York.